Evaluation of Automation Trustworthiness in Nuclear Power Plants: A Risk-Informed Approach using Probabilistic Validation and Integrated Probabilistic Risk Assessment

Muhammad Hammad Khalid[1*], Md Istiaque Ahmed[1], Samrendra Roy[2], Ha Bui[1], Seyed Reihani[1],
Ahmad Al Rashdan[3], Zahra Mohaghegh[1]
[1]Socio-Technical Risk Analysis (SoTeRiA) Laboratory
[2]Virtual Education and Research Laboratory (VERL)
[1,2]Department of Nuclear, Plasma, and Radiological Engineering, University of Illinois at Urbana-Champaign, 104 S. Wright Street, Urbana, IL 61801
[2]Department of Energy Science and Engineering, Indian Institute of Technology (IIT), Bombay, India 400076
[3]Idaho National Laboratory (INL), 1955 N Fremont Ave, Idaho Falls, ID 83415
*[riskanalysis@illinois.edu]

Abstract

The U.S. nuclear industry is progressively integrating automation technologies into Nuclear Power Plants (NPPs). To make informed decisions about large-scale investments in automation technologies specifically used for safety critical applications, stakeholders require robust evidence of their transparency, trustworthiness, and operational acceptability. This study introduces a risk-informed approach to evaluate automation trustworthiness by leveraging and making advancements to the Integrated Probabilistic Risk Assessment (I-PRA) methodological framework and the Probabilistic Validation (PV) methodology that were previously developed by some of the authors. I-PRA connects simulation models of underlying physical and social phenomena with the existing plant PRA model through a probabilistic interface. This study advances the I-PRA framework to explicitly capture relationships between the plant risk metrics and input parameters associated with the underlying human-automation-physics interactions. Meanwhile, the PV methodology is enhanced to characterize and propagate the uncertainties associated with the human-automation-physics coupling, allowing for the degree of automation trustworthiness to be measured by the magnitude of epistemic uncertainty associated with the automation output. Acceptability of a certain degree of automaton trustworthiness in a specific automation application is evaluated using predefined acceptance criteria that can be found available at one or more levels of the system hierarchy (e.g., automation output or plant risk). By integrating the advanced I-PRA framework with the enhanced PV methodology, the proposed approach offers a holistic evaluation of and an efficient algorithm to enhance automation trustworthiness. This paper also includes initial results of an ongoing case study evaluating the trustworthiness of an Artificial Intelligence (AI)-based automated firewatch system suggested for use in NPPs.

Keywords: Integrated Probabilistic Risk Assessment (I-PRA), Automation Trustworthiness, Probabilistic Validation (PV), AI-Based Automated Firewatch, Nuclear Power Plants.

1. Introduction

To improve efficiency and ensure safe, reliable operation, the U.S. nuclear industry is working to leverage automation as much as possible. Introduction of automation technologies, however, still presents challenging issues for most NPPs. These issues include defining an appropriate end-state automation architecture, developing business cases for automation implementation, assuring automation trustworthiness, improving automation transparency, and addressing licensing process burden. These challenging issues contribute to higher costs and schedule uncertainties for automation deployment, creating hurdles for the use of automation in NPPs. Before committing to a significant investment in the deployment of an automation technology in the nuclear domain, specifically for safety critical applications, decision-makers need methodologies that can generate sufficient evidence to verify that the automation would be explainable, trustworthy, and operationally acceptable. This paper focuses on establishing a theoretical basis and developing a generic methodology to evaluate automation trustworthiness.

In a companion paper [1], the authors have conducted a comprehensive literature review on the definition of automation trustworthiness and existing methodologies for its evaluation. Based on the results of this review, two common approaches for defining automation trustworthiness have been identified [1]: (i) Output deviation-based approach, i.e., defining automation trustworthiness using the deviation between the output of the automation model and the actual observations; and (ii) Attribute-based approach, i.e., defining automation trustworthiness using a set of attributes (e.g., “reliability and accuracy,” or “security, trust, resilience, and agility”). A methodology for evaluating trustworthiness associated with the first approach would require a substantial number of actual observations, which is not always the case in the nuclear context. On the other hand, using the second approach is subject to questions about, for example, the comprehensiveness of the attribute list and the appropriateness and adequacy of the attributes to be included. In the existing literature, evaluation of trustworthiness using this approach includes two aspects: (a) Measuring the attributes and (b) Aggregating the attributes. Measuring the attributes can be done either using expert opinions [2], data-driven approach [3], or model-based approach [4]. Meanwhile, aggregating the attributes can be done through either having a correlation that relates different attributes to the trustworthiness [5], or using a Bayesian Belief Network (BBN) [2].

To overcome the abovementioned limitations/challenges of the existing approaches, a systematic, risk-informed methodology for evaluating automation trustworthiness has been developed in this paper, using a third approach based on advanced uncertainty analysis. In this proposed methodology, automation trustworthiness is referred to as the “degree of confidence” that an automation system can function as expected; in this context, the “degree of confidence” can be measured by the magnitude of epistemic uncertainty associated with the automation system’s output. The proposed methodology leverages an Integrated Probabilistic Risk Assessment (I-PRA) methodological framework and a Probabilistic Validation (PV) methodology previously developed by some of the authors and makes necessary advancements to enable the evaluation of automation trustworthiness in NPP context.

I-PRA was developed by some of the authors in their previous studies [6, 7] to allow for integrating simulation models of underlying physical and social phenomena with the existing plant PRA model through a probabilistic interface. In the current study, I-PRA is advanced by adding and/or advancing automation-related modules (i.e., those red-boxed modules in Figure 1) to enable a clear tracing and capture of relationships between the plant risk metrics and input parameters associated with the underlying physics, automation, and human performance.

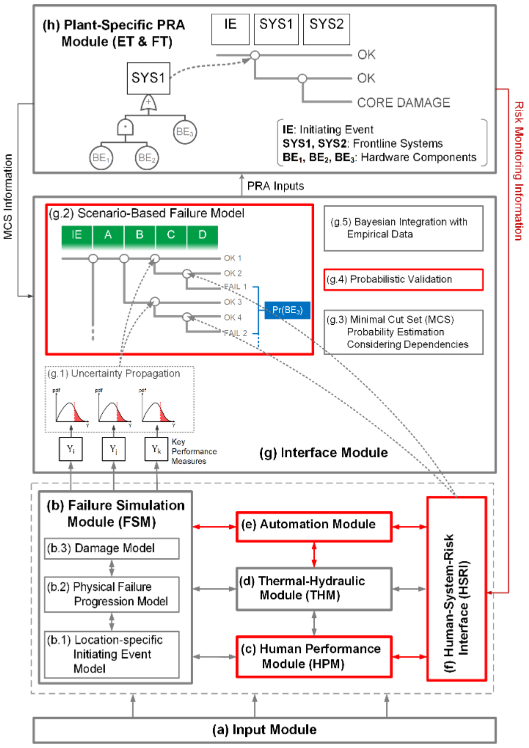
(h) Plant-Specific PRA Module (ET & FT)
IE
SYS1
SYS2
OK
OK
CORE DAMAGE
SYS1
BE₁
BE₂
BE₃
IE: Initiating Event
SYS1, SYS2: Frontline Systems
BE₁, BE₂, BE₃: Hardware Components
PRA Inputs
MCS Information
Risk Monitoring Information
(g.2) Scenario-Based Failure Model
IE
A
B
C
D
OK 1
OK 2
FAIL 1
OK 3
OK 4
FAIL 2
Pr(BE₃)
(g.5) Bayesian Integration with Empirical Data
(g.4) Probabilistic Validation
(g.3) Minimal Cut Set (MCS) Probability Estimation Considering Dependencies
(g.1) Uncertainty Propagation
pdf
Y
Yᵢ
Yⱼ
Yₖ
Key Performance Measures
(g) Interface Module
(b) Failure Simulation Module (FSM)
(b.3) Damage Model
(b.2) Physical Failure Progression Model
(b.1) Location-specific Initiating Event Model
(e) Automation Module
(d) Thermal-Hydraulic Module (THM)
(c) Human Performance Module (HPM)
(f) Human-System-Risk Interface (HSRI)
(a) Input Module

Figure 1: Integrated PRA (I-PRA) Framework for Automation Technology

The PV methodology was originally developed to test the validity of simulation models [8, 9] and is leveraged and advanced in the current study for evaluating automation trustworthiness. Following this advancement, PV allows for characterizing and propagating uncertainties associated with the human-automation-physics coupling and quantifying the magnitude of epistemic uncertainty associated with the automation output, a.k.a., the degree of automation trustworthiness. Execution of the PV methodology for evaluation of automation trustworthiness in NPPs requires that the uncertainties from the underlying automation model level be propagated up to the level of plant risk metrics, which necessitates a computational platform that is achievable by leveraging the advanced I-PRA framework. The acceptability of the degree of automaton trustworthiness for a specific automation application is evaluated against predefined acceptance criteria at the output level or plant risk metric level. The PV methodology is also equipped with an advanced Global Importance Measure (GIM) method to rank significant sources of uncertainty at the level of underlying physics, automation, and human performance concerning their contribution to the plant risk uncertainty. This provides an efficient algorithm to enhance automation trustworthiness.

To demonstrate the feasibility and practicality of the advanced I-PRA framework and the PV methodology, this paper also reports on results of an ongoing case study of AI-based automated firewatch system. By integrating the advanced I-PRA framework with the PV methodology, the proposed approach offers a holistic evaluation of automation trustworthiness, considering complex interactions among human operators, automation technologies, and the evolution of physical phenomena in NPPs. This comprehensive evaluation will assist in making informed decisions about the deployment and operation of automation technologies in NPPs, ultimately contributing to the enhancement of plant safety and efficiency.

The remainder of this paper is organized as follows. Section 2 explains the proposed methodology for evaluating automation trustworthiness. Section 3 presents an application of the developed methodologies for a case study that centers around an AI-based automated firewatch system. Finally, Section 4 provides concluding remarks and discusses future work.

2. Methodology to Evaluate Automation Trustworthiness

As asserted, the proposed methodology for automation trustworthiness leverages the I-PRA methodological framework and the PV methodology previously developed by some of the authors. Necessary advancements to I-PRA and PV are made in the proposed methodology to enable the evaluation of automation trustworthiness in the NPP context. These advancements to the I-PRA framework and PV methodology are discussed in Subsections 2.1 and 2.2, respectively.

2.1. Advancing the Integrated Probabilistic Risk Assessment (I-PRA) Methodological Framework

The fundamental concept of the I-PRA methodological framework [6, 7] is that simulation models of underlying physical and social phenomena are developed and then integrated with the existing plant PRA through a probabilistic interface. I-PRA adds realism into the risk estimation by explicitly incorporating time and space into underlying models while avoiding significant changes to the plant PRA model and its associated costs (e.g., peer review). In previous work [6, 7, 10], I-PRA encompasses Failure Simulation Module (FSM) ('b' in Figure 1), Human Performance Module (HPM) ('c' in Figure 1) and their couplings. The dynamic coupling between the modules is similar in its nature to simulation-based PRA (or dynamic PRA), while the Interface Module ('g' in Figure 1) conducts uncertainty propagation and offers the possibility of using simulation approaches linked with the existing PRA of plants ('h' in Figure 1).

In the current regulatory and nuclear industry PRA practices, simulations are used in various applications to estimate PRA inputs. For instance, Fire PRA uses a fire model, e.g., CFAST [11], to estimate cable damage probabilities. The connection of these simulation models with PRA, however, is 'passive,' where the underlying simulation generates basic event probabilities as its outputs, which are then plugged into the PRA software as inputs [12]. This passive PRA-simulation interface does not have the capability to trace the input-output relationships between plant risk metrics and the input parameters of the simulation models. In contrast, I-PRA creates a "unified" connection between the plant PRA and the underlying physics and human performance simulations, where the transfer of data and information among multiple levels of causality (physical mechanisms, human performance, and plant-level PRA) are carried out in a cohesive computational platform. This way, the relationship between the plant risk metrics and the input parameters associated with the underlying physics and human performance can be explicitly traced and captured. The unified connection also allows for adding a GIM method to I-PRA to enable the ranking of significant sources of uncertainty at the level of underlying physics and human performance with respect to their contribution to the uncertainty in the plant risk. I-PRA has been successfully used in two industry applications: (i) the risk-informed resolution of Generic Letter 2004-02 [7, 13], and (ii) Fire PRA [6, 10, 14].

To appropriately credit the use of automation technologies in the PRA of NPPs, this current work advances the existing I-PRA framework by (i) adding the Automation Module (AM) ('e' in Figure 1) and the Human-System-Risk Interface (HSRI) ('f' in Figure 1); (ii) advancing the HPM and the Scenario-Based Failure Model ('g.2' in Figure 1); and (iii) developing a computational coupling among those added modules and the existing human and system performance modules in I-PRA. The upgraded I-PRA framework simulates complex interactions among automation systems, physical phenomena, and plant operators. It allows for the impacts of these interactions on plant equipment and system performance to be evaluated. The following sub-sections discuss the advanced I-PRA framework and its automation-related modules. Since our case study involves an AI-based automated firewatch system that can be credited in Fire PRA, the discussion will use this case study and the Fire PRA context to explain details of the automation-related modules and the spatiotemporal interactions among them.

2.1.1. Advanced I-PRA Framework for the AI-based Automated Firewatch Case Study

When developing the I-PRA framework for the case study of AI-based automated firewatch at an NPP, the first task was to work out the interactions of interest (i.e., related to the automated firewatch) among the AM, HPM, and FSM. Then based on the identified interactions, these modules were developed in detail and coupled with each other inside the I-PRA framework, shown in Figure 2 and explained in the following subsections. This coupled environment is in essence the lower portion of the I-PRA framework with modules in them in Figure 1. While Figure 1 depicts the interactions among models in a more general sense, Figure 2 offers a detailed resolution in terms of this study.

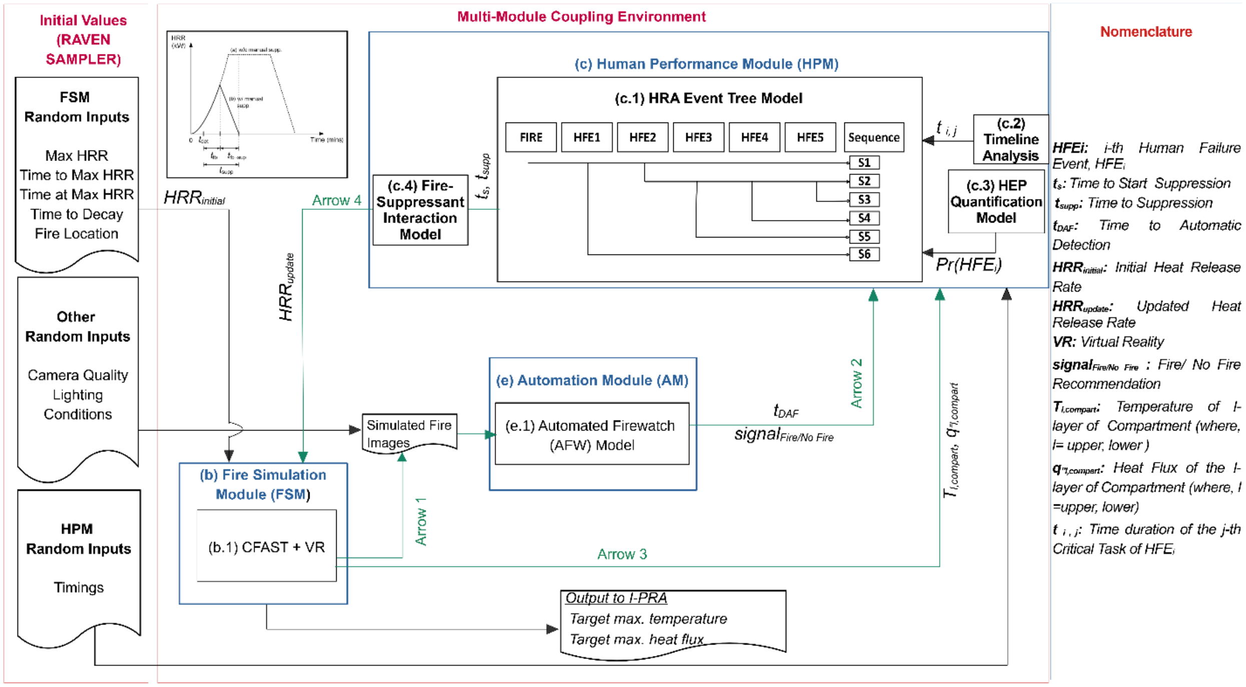
Initial Values (RAVEN SAMPLER)
FSM Random Inputs
Max HRR
Time to Max HRR
Time at Max HRR
Time to Decay
Fire Location
Other Random Inputs
Camera Quality
Lighting
Conditions
HPM Random Inputs
Timings
HRR (kW)
(a) w/o manual supp.
(b) w/ manual supp.
Time (mins)
0
t_{det}
t_{fb}
$t_{fb-supp}$
t_{supp}
$HRR_{initial}$
HRR_{update}
Multi-Module Coupling Environment
(c) Human Performance Module (HPM)
(c.1) HRA Event Tree Model
FIRE
HFE1
HFE2
HFE3
HFE4
HFE5
Sequence
S1
S2
S3
S4
S5
S6
(c.2) Timeline Analysis
(c.3) HEP Quantification Model
(c.4) Fire-Suppressant Interaction Model
t_s, t_{supp}
$t_{i,j}$
$Pr(HFE_i)$
Arrow 4
Arrow 2
Arrow 1
Arrow 3
Simulated Fire Images
(e) Automation Module (AM)
(e.1) Automated Firewatch (AFW) Model
t_{DAF}
$signal_{Fire/No\ Fire}$
$T_{l,compart}, q''_{l,compart}$
(b) Fire Simulation Module (FSM)
(b.1) CFAST + VR
Output to I-PRA
Target max. temperature
Target max. heat flux
Nomenclature
$HFEi$: i-th Human Failure Event, HFE_i
t_s: Time to Start Suppression
t_{supp}: Time to Suppression
t_{DAF}: Time to Automatic Detection
$HRR_{initial}$: Initial Heat Release Rate
HRR_{update}: Updated Heat Release Rate
VR: Virtual Reality
$signal_{Fire/No\ Fire}$: Fire/ No Fire Recommendation
$T_{l,compart}$: Temperature of l-layer of Compartment (where, l= upper, lower)
$q''_{l,compart}$: Heat Flux of the l-layer of Compartment (where, l =upper, lower)
$t_{i,j}$: Time duration of the j-th Critical Task of HFE_i

Figure 2: Coupling Methodology for Automation, Human Performance, and Fire Simulation Module

2.1.1.1. <u>Fire Simulation Module (FSM)</u>

FSM simulates the initiation and propagation of the physics of fire. The Consolidated Fire and Smoke Transport (CFAST) software and Virtual Reality (VR) tool are used for developing the FSM. CFAST is a fire model that can create and predict fire initiation and progression environment inside a compartment. It is a two-zone fire model that divides the compartment into an upper layer and lower layer during analysis and helps in calculating the time-evolving distributions of gas temperature inside the layers, temperatures at pre-defined targets as well as concentrations of smoke and other gaseous byproducts created during the combustion process inside the compartment. In the previous work, CFAST [10] and Fire Dynamics Simulator (FDS) [14] are used to simulate the physics of fire. FDS as another fire and smoke model, can perform simulations at much higher resolution. For the present case study, CFAST is used for simplicity and the outputs of CFAST are illustrated in a virtual environment using the Unity software [15]. VR has been used as a substitute to experiments and it is facilitating the present case study by providing: (i) high quality fire images to train and validate AI-based fire classifier; and (ii) more control over the environment subjected to fire. This would help in analyzing the influence of a large number of parameters such as lightening conditions, camera quality and camera position over AI-based fire classifier's performance. This is important because eventually while estimating the trustworthiness of automation system using uncertainty-based approach, identification and characterization of underlying uncertain parameters would be needed.

It should be noted that training and validating the AI-based model on the data coming actual experiments should always be preferred over the data generated by VR. However, availability of data coming from the real world that matches well with the actual scenarios of interest could be challenging at times. Using available but irrelevant real world fire images to train or test AI-based classifier will not be the best way moving forward as it halts the synergy between fire initiation and progression parameters of interest such as Heat Release Rate (HRR) curves, that were used to run CFAST model. Hence, to maintain that synergy, fire scenarios illustrated inside VR (Consequently fire images to train/test AI-based fire classifier) were made consistent with the ones created inside CFAST environment. This is done by creating fire scenarios inside VR using the outputs of CFAST model such as time evolution of HRR and soot concentrations.

2.1.1.2. Automation Module (AM)

The Automation Module (AM) ('e' in Figure 2) contains an Automated Firewatch (AFW) model ('e.1' in Figure 2), which is an AI-based fire classifier and is developed using a Convolutional Neural Network (CNN). For the present case study, the fire classifier is binary (Fire/no-fire) in nature. Sigmoid and Rectified Linear Unit activation function (ReLU) are used to add the desired level of non-linearity, which is needed to model complex functions, enhancing the capabilities of the developed classifier. In order to select the most optimum parameters during the training phase of the classifier, insights regarding performance of the classifier at various sets of parameters are needed, that would be coming from estimating the loss function of the classifier. In this work, binary cross entropy is used to compute the desired loss function, that is one of the most common loss functions used to train binary classifiers. The AM model architecture in this paper has been illustrated in Figure 3. In step 1, a rescaling layer normalizes the pixel values of the imagery data, reducing the computational load and making the training process more stable. Then, step 2 enhances the dominant features of the images such as edges and reduces image array dimensions, to improve model efficiency and helps in noise reduction. Finally, in step 3, a multidimensional output (from convolutional or pooling layers) is converted to one dimensional vector, which can be fed into the subsequent fully connected layers. The final layer is coupled with a Sigmoid function that creates the boundary between the two classes (fire/no-fire) for the images fed into the architecture.

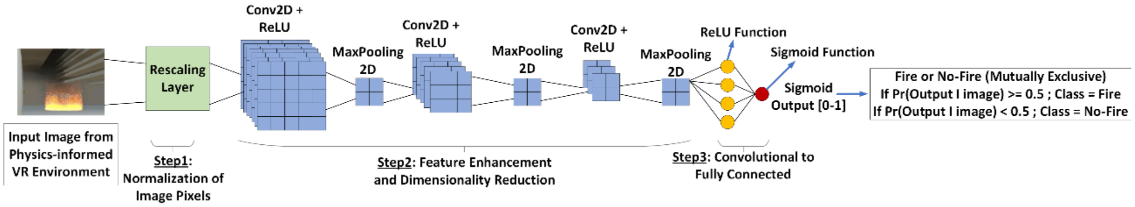
Input Image from Physics-informed VR Environment
Rescaling Layer
Conv2D + ReLU
MaxPooling 2D
Conv2D + ReLU
MaxPooling 2D
Conv2D + ReLU
MaxPooling 2D
ReLU Function
Sigmoid Function
Sigmoid Output [0-1]
Fire or No-Fire (Mutually Exclusive)
If Pr(Output I image) >= 0.5 ; Class = Fire
If Pr(Output I image) < 0.5 ; Class = No-Fire
Step1: Normalization of Image Pixels
Step2: Feature Enhancement and Dimensionality Reduction
Step3: Convolutional to Fully Connected

Figure 3: Model Architecture of AI-Based Fire Classifier Using CNN

In this paper, the training of the AI-based fire classifier followed the pattern of physics-informed machine learning. At first, fire simulation module using CFAST has been run for different fire initiation and propagation scenarios (each sampled set of inputs forms a specific scenario) that were visualized in Virtual Reality (VR) environment [11]. Then, the captured images are annotated with their desired classes as fire or no-fire using the same CFAST data based on actual physics. This is done in a way that an image simulating a scenario in VR environment at an instant where the HRR is non-zero/zero has been labeled as a fire/no-fire image. Training has been performed on 50 images from each of the 200 different scenarios (generated a total of 10000 images) representing a sampled set of all the input features, such as HRR curve of the fire and environmental conditions in VR. A training accuracy of around 99% and a validation accuracy of around 98.5% have been achieved by the end of 8 epochs.

2.1.1.3. Human Performance Module (HPM)

This module is developed using the IDHEAS HRA methodology. In this study, the HPM is advanced (as compared to previous Fire I-PRA studies by Sakurahara et. el. [14])) by extending its scope to include both Main Control Room (MCR) operator actions and first responder (FR) actions (outside the MCR) in response to the automation signals provided through the HSRI. There are multiple human failure events (HFEs) in the HRA based Event Tree Model, e.g., HFE_1 (Operator Confirms and Communicate with FR), representing the human action of the operator, where, given a cue from the AM, the operator is supposed to communicate with a FR (another human agent) and dispatch the agent to the site of the fire signal to confirm the occurrence of a fire event. Furthermore, the time associated with each of the HFEs is also obtained from the HPM and those time durations are utilized to calculate the time to start suppression, the details of which can be found in subsection (c.2).

(c.1) HRA Event Tree Model

There are multiple HFEs, e.g., HFE_1, HFE_2, representing the human failure events of the HRA Event Tree Model as shown in Figure 4. There are also two end states, OK and NS. OK end states represent the sequence where suppression event is successful. On the other hand, NS denotes the sequence where neither of the suppression top events are successful. i.e., the fire is not suppressed.

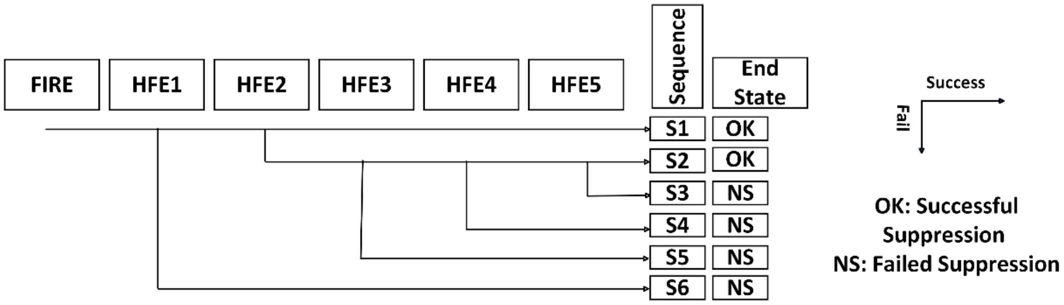

Figure 4: HRA Event Tree Model for HPM (c.1)

The HEPs for those top events were calculated from the HEP Quantification Model (c.3) and utilized in the coupled simulations to determine the number of times these top HFEs should be assumed to succeed and fail. The HFEs are modeled in such a way that there is only one end state for the model of each HFE. Modeling an HFE this way allows for calculating the HEP for the whole HFEs using IDHEAS. In this case, since there is only one success state, it means that for the HFE to be successful, all involved tasks must be successful i.e., the number of critical tasks for the analysis of the HFE using IDHEAS is equal to the total number of tasks. Note that, since there is a separate Timeline Analysis Model being done in parallel, the HEP is estimated using IDHEAS only accounting for the Cognitive Failure Modes (CFM) i.e., for each analysis of the HFE, the probability of failure due to insufficient time (P_t) is set to be zero. The damage to a target due to human failure because of insufficient time is reflected on the modified HRR curve i.e., the target cable damaged because there was not enough time for the human agents to successfully perform an HFE, will be reflected through the HRR curve being modified at a later point in time resulting in the max temperature exceeds the melting point of the cable and thereby damaging the cable. The list of HFEs and their constituent Critical Tasks (CTs) in addition to the duration of the CTs are described in Table 1.

The first HFE_1 is the one where the operator confirms and communicates with the FR. The suppression by the FR top event, $HFE_{2,}$ consists of three human tasks and has only one success state. The HFE_2 is influenced by the environmental conditions since the assumption here is that the first and the second Critical Tasks (CTs) are dependent on the environmental conditions. Based on the analysis here, there are two relevant Performance Shaping Factors (PSFs), the first is related to the temperature of the compartment and the other one which is related to the visibility of the compartment. Visibility is calculated using the correlation described in the works of Bui et. el. [2] (see c.2 Timeline Analysis Model). The PSFs are included for a critical task in the calculation if at the start of that critical task either the upper layer or the lower layer temperature exceeds a specific threshold and visibility falls below a threshold (in this case, the threshold for temperature is 33^0C and the threshold for visibility is 40m, the highest dimension of the cable tray room). The HFE_3 consists of tasks that are outside the location of fire and hence not influenced by fire induced environmental conditions. This HFE_4, Fire Brigade Activity 1, consists of tasks performed by the fire brigade team immediately before the start of suppression. Finally, the HFE_5 consists of tasks involving the suppression activity of fire brigade. This suppression by the fire brigade team consists of tasks that are executed at the location of fire. Like HFE_2, the PSFs are included in the calculation if at the start of each critical task either the upper layer or the lower layer temperature and visibility exceeds specific thresholds. An additional assumption is that, if a PSF is selected for one of the Critical Tasks then the PSF is included for all the CFMs (e.g., Understanding, Action) inside that critical task e.g., if the temperature PSF is selected.

Table 1: Human Failure Events (HFEs) and Constituent Critical Tasks (CTs)

HFE_i	**Critical Tasks**	**Critical Task ID ($CT_{i,j}$)**	**Cognitive Failure Modes (CFM)**	**Total Duration ($t_{i,j}$)**
HFE_1	Operator Receives Fire Detection from AM and Diagnoses the Signal on the MCR Computer	$CT_{1,1}$	Detection, Understanding, Deciding	$t_{1,1}$
	Operator Communicates with First Responder (FR)	$CT_{1,2}$	Action	$t_{1,2}$
	FR Arrives at and enters Fire Location, and Communicates with MCR Operator [16]	$CT_{1,3}$	Understanding, Deciding, Action	$t_{1,3}$
HFE_2	FR Searches for and Locates Fire Source	$CT_{2,1}$	Detection, Action	$t_{2,1}$
	FR Starts manually Suppressing Fire using Portable Fire Extinguisher	$CT_{2,2}$	Action	$t_{2,2}$
HFE_3	Operator Receives Signal from FR to Call the Fire Brigade (FB)	$CT_{3,1}$	Detection, Understanding, Deciding	$t_{3,1}$
	Operator Communicates with Fire Brigade	$CT_{3,2}$	Action	$t_{3,2}$
HFE_4	FB Gathers at the Assembly Area and Gets Pre Plan	$CT_{4,1}$	Action	$t_{4,1}$
	FB Peer Checks and Moves to and Arrives at the Staging Area	$CT_{4,2}$	Action, Inter Team	$t_{4,2}$
	FB Lays Hose & Puts on Breathing Apparatus	$CT_{4,3}$	Action	$t_{4,3}$
HFE_5	FB Searches for and Locate Fire	$CT_{5,1}$	Detection, Action	$t_{5,1}$
	FB Prepares Hose for Discharging and Starts Suppressing the Fire	$CT_{5,2}$	Action	$t_{5,2}$

(c.2) Timeline Analysis

The objective of the timeline analysis (Figure 5) is to keep track of the duration of the Critical Tasks (CT) and calculate the time to start suppression for the Fire Suppressant Interaction Model (c.4). The constituents of Critical Tasks of the HFEs and their time durations along with the variable used to denote the time durations of these critical tasks are described in Table 1. These time durations are sampled from the distributions described in Table 3 and are used to obtain the time to start suppression. After obtaining the time to start suppression from this step, the HRA Event Tree Model is used along with the HEP quantification model (c.3) for a specific run of the coupled code to generate an updated HRR curve that refed to FSM to get the KPMs of interest. The consecution of these critical tasks can be seen in Figure 5.

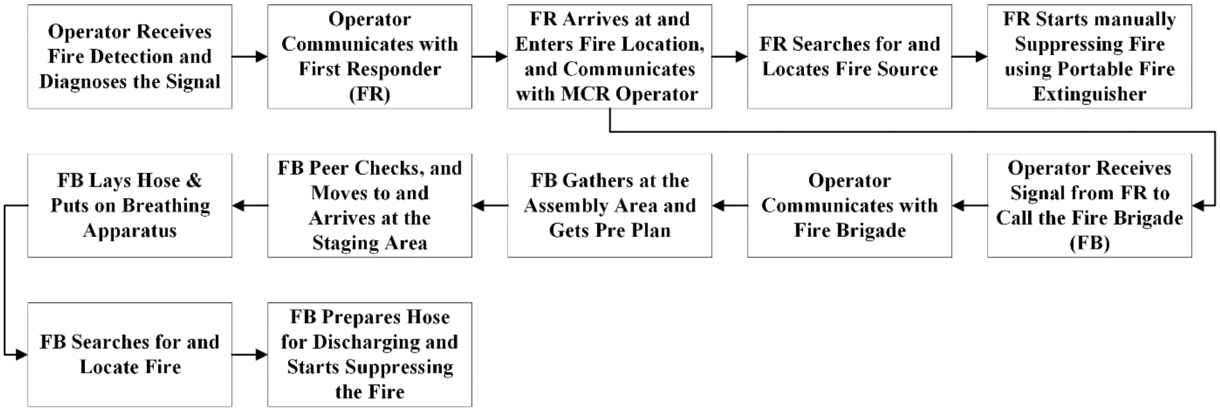
Operator Receives Fire Detection and Diagnoses the Signal
Operator Communicates with First Responder (FR)
FR Arrives at and Enters Fire Location, and Communicates with MCR Operator
FR Searches for and Locates Fire Source
FR Starts manually Suppressing Fire using Portable Fire Extinguisher
Operator Receives Signal from FR to Call the Fire Brigade (FB)
Operator Communicates with Fire Brigade
FB Gathers at the Assembly Area and Gets Pre Plan
FB Peer Checks, and Moves to and Arrives at the Staging Area
FB Lays Hose & Puts on Breathing Apparatus
FB Searches for and Locate Fire
FB Prepares Hose for Discharging and Starts Suppressing the Fire

Figure 5: Timeline of Critical Tasks for the HFEs in HRA Event Tree Model (c.1)

. To determine the time to start fire suppression based on different scenarios a sampling technique was adopted from the paper by Sakurahara et. al. [20]. This method allows for controlling the simulation by adopting the split fraction method. By random sampling a number from a uniform distribution and comparing with the probability, the number of simulations is being split according to the ratio of event's states, in this case, success and failure. The timeline analysis can be divided into the following steps.

Step 1: Start by checking a random value between 0 and 1 and compare it against the probability of Detection by the Automated Firewatch, Pr(DAF). If this random value is greater than Pr(DAF), proceed to step 2.a, otherwise proceed to step 2.b. This Pr(DAF) value will be provided from the AM.

Step 2.a: The next check will involve another random value against Pr(HFE_1) (from HEP quantification model). If the random number is greater, proceed to step 3. Otherwise, the HRR curve is not suppressed.

Step 3: Get the upper- and lower-layer temperature ($T^k_{l,compart}$ (t); $l =$ upper, lower) , and optical density ($OD^k_{l,compart}$ (t); $l =$ upper, lower) values of the compartment, from the non-suppression HRR curve output at,

$$t_1 = \sum_{j=1}^{3} t_{1,j} \; ; \qquad \text{(Eq. 1)}$$

which is the start of searching for and locating fire source, i.e., the first critical task of HFE_3. In this study we calculated the visibility data from the empirical correlation used in the works of Bui et. al. [10]. Here, visibility,

$$v = k_v \frac{1}{OD_{l,compart}(t_1);}; l = \text{upper, lower}; \qquad \text{(Eq. 2)}$$

Now set the environmental PSFs for $CT_{2,1}$ of HFE_2, if any of the data exceeds some threshold value set by the expert. Then set the PSFs for $CT_{2,2}$ of HFE_2 by getting the environmental conditions of the compartment at,

$$t_1 + t_{2,1}; \qquad \text{(Eq. 3)}$$

Now use HEP quantification model to calculate the probability of HFE_2 Pick a random number from a uniform distribution from 0 to 1. If this random number is greater than Pr(HFE_2), proceed to step 4.a. Otherwise proceed to step 4.b.

Step 4.a: The top event HFE_2 is successful. The time to start suppression,

$$t_s = t_1 + t_{2,1}; \qquad \text{(Eq. 4)}$$

Step 4.b: From HEP quantification model, we get the Pr(HFE_3) which is the probability of failure for the Pre Fire Brigade Activity. Pick a random number from the uniform distribution. If it is greater than, Pr(HFE_3), go to step 5.a. Otherwise, go to step 5.b.

Step 5.a: Pick a random number from a uniform distribution from 0 to 1. If it is greater than the Pr(HFE_4) which is the human error probability of failure for the top event Fire Brigade Activity 1, go to step 6.a. Otherwise, go to step 6.b.

Step 6.a: Now, the top event HFE_5, is performed inside the fire room. Check the environmental conditions (i.e., the upper layer or lower layer temperatures and optical densities of compartment) at,

$$t_2 = \sum_{j=1}^{3} t_{1,j} + \sum_{j=1}^{2} t_{2,j} + \sum_{j=1}^{2} t_{3,j} + \sum_{j=1}^{3} t_{4,j}; \qquad \text{(Eq. 5)}$$

$$t_{i,j} = \text{time duration of } \mathrm{CT_{i,j}} \text{ for } \mathrm{HFE_i}$$

And at

$$t_2 + t_{5,1};$$
$$t_{5,1} = \text{time duration of } \mathrm{CT}_{5,1} \text{ for } \mathrm{HFE}_5; \qquad \text{(Eq. 6)}$$

Get the Pr(HFE_5) by setting the appropriate PSFs for CT_1 and CT_2. If the random number exceeds Pr(HFE_5), then HFE_5 is a success and $\mathrm{t_s} = \mathrm{t_2} + \mathrm{t_{5,1}}$.Otherwise, the HRR curve is non-suppressed.

Step 6.b: If the random number is smaller than Pr (HFE_4), then the HRR curve is non suppressed.

Step 5.b: If the random number is smaller than Pr(HFE_3), the HRR curve is non suppressed.

Step 2.b: In this scenario DAF is a failure. Therefore, HRR curve is non suppressed.

(c.3) HEP Quantification Model: To quantify the failure probabilities of the HFEs IDHEAS [23] software developed by the NRC is used. This is because of the widespread acceptability of the use of IDHEAS in the quantification process of the Human Error Probability.

(c.4) Fire Suppressant Interaction Model:

After obtaining the time to start suppression, t_s , the updated HRR curve, $\dot{\mathrm{Q}}_\mathrm{s}(\mathrm{t})$, is generated and used to rerun the FSM. To generate $\dot{\mathrm{Q}}_\mathrm{s}(\mathrm{t})$, a correlation was used in the works of Bui et. el. [10] and it has been adopted in this study.
One of the outputs of this model is the time to full suppression,

$$t_{FS} = t_s + \frac{1}{k_f} \ln\left[\frac{\dot{Q}_{controlled}}{\dot{Q}_0(t = t_s)}\right] \qquad \text{(Eq. 8)}$$

Here,
$\dot{Q}_{controlled}$ = Heat Release Rate when fire is considered suppressed
$\dot{Q}_0(t = t_s)$ = Heat Release Rate at the start of suppression, t_s
k_f = coeffient for calculating the time to full suppression

Yu et al. [24] derived that, under the assumptions that the fuel material properties are constant over temperature and that the water evaporation rate per unit burning surface area is always balanced with the water application rate per unit area (denoted by m''_w), the coefficient k_f can be formulated as follows:

$$k_f = a_1 m''_w + a_2 \qquad \text{(Eq. 9)}$$

where a_1 and a_2 are represented by analytical functions of several physical parameters. The empirical correlation developed for the FMRC standard plastic commodity suggests that $a_1 = 0.716$ and $a_2 = 0.1317$ [24].
Using this time to suppression and time to start suppression from the timeline analysis model, the HRR curve can be updated to reflect the suppression event.

2.1.1.4. *Identified interactions among the AM, HPM, and FSM for the Automated Firewatch case study.*

These interactions are represented by six arrows on Figure 2, numbered from 1 to 6, and are explained below.

- Arrow 1 (FSM to AM): The FSM models the physics of fire and will provide simulated fire images for the AM.
- Arrow 2 (AM to HPM): The AM can provide information related to the time evolution of fire and thereby reduce the time it takes to detect fire. This in turn, will affect the time available and, by proxy, the performance of the responders.
- Arrow 3 (FSM to HPM): The FSM provides data associated with the physics of fire, e.g., temperature, smoke density, heat flux etc. These physical parameters can influence the performance of the human. For example, reduced visibility due to smoke can affect the mobility of humans and thereby affect performance.
- Arrow 4 (HPM to FSM): The effect of human performance can be reflected on the propagation of the fire. Its impact can be inferred by the effect of human performance on the HRR curve which is an input for the FSM. Here, the effect is reflected by updating the HRR curve based on the result of the HPM.

2.1.1.5. *Scenario-Based Failure Model*

Once the FSM, AM, HPM are developed, they are coupled together and are connected to a Scenario-Based Failure Model ('g.2' in Figure 1) and the plant PRA Module ('h' in Figure 1) using a computational platform. Details of the couplings are discussed in Section 2.1.2. The Scenario-Based Failure Model in this context was developed by leveraging the Fire Detection and Suppression Event Tree in Appendix P of the NUREG/CR-6850. In that Event Tree, three types of fire detection and suppression are considered: prompt, automatic, and manual. The Scenario-Based Failure Model maintains these types and considers the Automated Firewatch System as an additional measure to be included for "Prompt" Detection (instead of Prompt detection by continuous human firewatch and/or continuously occupied personnel in NUREG/CR-6850). The Scenario-Based Failure Model has multiple sequences (depending on the status of the Prompt/Automatic/Manual Detection and Suppression functions), and these sequences are connected to different end states that are associated with fire-induced damage to target components.

The Scenario-Based Failure Model is then connected to the plant-specific PRA Module ('h' in Figure 1) to credit the automated firewatch in the Fire I-PRA framework. The automated firewatch case study evaluates the effectiveness of the AI-based automated firewatch in a fire scenario and analyzes its impacts on plant risk metrics (i.e., core damage frequency [CDF]).

2.1.2. FSM-AM-HPM Couplings for the AI-based Automated Firewatch Case Study

This case study uses an NPP fire scenario from NUREG-1934 Scenario E [17]. This scenario estimates the impact of a transient fire in a trash receptacle on the cables placed in cable trays inside a Cable Spreading Room (CSR) as shown in Figure 6.

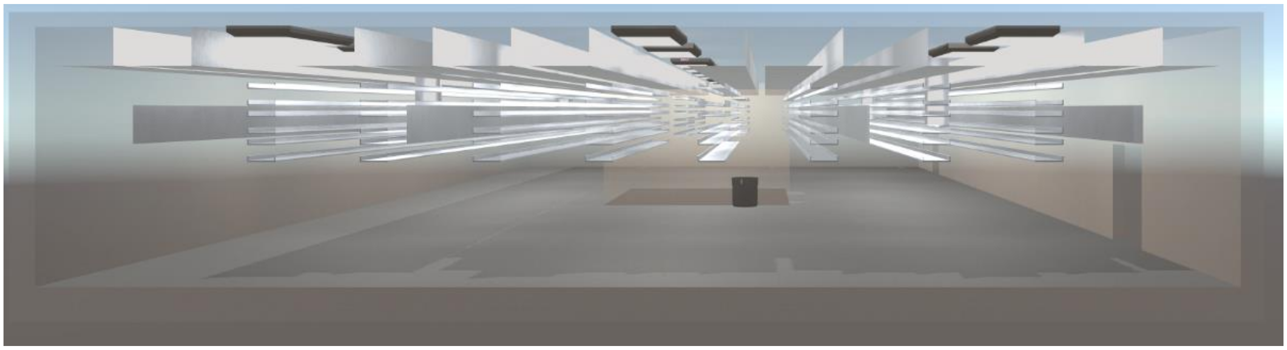

Figure 6: Configuration of Switch Gear Room (SGR) used in the case study (NUREG-1934 Scenario E)

The modeling methodology of coupling between the FSM, HPM and AM for an internal fire scenario is shown in Figure 2. The I-PRA framework (Figure 1) provides a means of accomplishing that objective while considering the dependencies among the modules. The coupled simulation environment provides a means to treat the dependencies that are otherwise implicit by putting all the interactions in the coupled environment. For a given set of epistemic parameters, the process for performing quantitative analysis is broken into multiple steps. Let $g_{FSM}(\cdot)$ be the functional representations of the FSM model and t_{BO} be the time when the fire self-extinguishes under the burnout condition (without any fire protection features) while $g_{HPM}(\cdot)$ is the functional representations of human performance model. Under this setting, the "discrete" FSM-Automation-HPM coupling is developed as follows:

<u>*Step 1. Random Sampling.*</u>

Once the epistemic parameters are fixed, create random samples for the uncertain parameters and variables, particularly the time distributions of each of the top events and the random variables associated with the HRR curve. Each set of random samples for parameters of the HRR represents the unsuppressed curve. For Random sampling RAVEN [26] is used because it offers extended ability for various uncertainty calculation processes.

Vector of Random variables for the aleatory parameters are generated.

$$\mathrm{X} = (X_{FSM}, X_{HPM}, X_{AM}); \qquad \text{(Eq. 10)}$$

$$\text{Here, } X_{FSM} = input\ vector\ for\ FSM,$$
$$X_{HPM} = input\ vector\ for\ HPM,$$
$$X_{AM} = input\ vector\ for\ AM,$$

<u>*Step 2. Running FSM with unsuppressed HRR Curve*</u>

Run the FSM Model with the unsuppressed HRR curve, aimed at finding the temperature of a specific compartment and specific targets situated in the compartment.

$$T(t)_{l,compart} = \mathrm{g_{FSM}}(Q_0(X_{FSM}, \mathrm{t}), X_{FSM})\ ;\ 0 \leq \mathrm{t} \leq \mathrm{t_{BO}}, \qquad \text{(Eq. 11)}$$

$$OD(t)_{l,compart} = \mathrm{g_{FSM}}(Q_0(X_{FSM}, \mathrm{t}), X_{FSM})\ ;\ 0 \leq \mathrm{t} \leq \mathrm{t_{BO}}, \qquad \text{(Eq. 12)}$$

$$T(t)_{i,target} = \mathrm{g_{FSM}}(Q_0(X_{FSM}, \mathrm{t}), X_{FSM})\ ;\ 0 \leq \mathrm{t} \leq \mathrm{t_{BO}}, \qquad \text{(Eq. 13)}$$

where $T(t)_{l,compart}$and $OD(t)_{l,compart}$ are the temperature and the optical density of the layer l (upper/lower) at time t of the compartment, respectively, $T(t)_{i,target}$ is the temperature of the i^{th} (maximum value of i is the total number of targets) target in the compartment which can be any of the physical components of the automation module as well, $Q_0(\bullet)$ is the burnout HRR curve without suppression, X_{FSM} is a vector of input parameters for FSM, some of which are (e.g., max HRR, time to max HRR etc.) used for constructing the burnout HRR curve for a sample set.

Step 3. Obtain Time to Detection by Automated Firewatch, $t_{DAF}$from AM and Perform Timeline Analysis

Create simulated fire images of the fire scenario until the time of starting suppression using the virtual reality (VR) from the values of parameters of the HRR curve. The output AM, in this case is the time to detection by automated firewatch, t_{DAF}. It can be functionally represented as follows,

$$\mathrm{g_{AM}}\big(OD(t)_{l,compart}, Q_0(X_{FSM}, \mathrm{t}), X_{AM}\big)\ ;\ 0 \le \mathrm{t} \le \mathrm{t_{BO}}, \quad \text{(Eq. 14)}$$

Where $OD(t)_{l,compart}$ is one of the outputs of the FSM generated by the non-suppression HRR curve.

$$OD(t)_{l,compart} = \mathrm{g_{FSM}}(Q_0(X_{FSM}, \mathrm{t}), X_{FSM})\ ;\ 0 \le \mathrm{t} \le \mathrm{t_{BO,}} \quad \text{(Eq. 15)}$$

Timeline analysis model needs to be performed here to obtain the time to suppression, t_s which requires the time to detection by the Automated Firewatch Model (sub section 2.1.1.3 subsection c.2). This whole process of obtaining t_s can be functionally represented as an output of the HPM. After determining which scenario is being analyzed in the coupled simulation, the time to start suppression can be derived from the HPM. If in the scenario the fire is not suppressed, then the time to start suppression is set to time to burnout ($\mathrm{t_{BO}}$).

$$\mathrm{t}_s = \mathrm{g_{HPM}}(\, X_{HPM}) \quad \text{(Eq. 16)}$$

Then we have a cross pairs of time to start suppression and their updated HRR curve that will be determined by the fire-suppressant interaction model (c.4)

$$\left[\mathrm{t}_s, Q(t)_{update}\right] \quad \text{(Eq. 17)}$$

Step 4: Rerun CFAST using the Updated HRR Curve

Rerun the FSM Model using $\dot{\mathrm{Q}}_{\mathrm{Mod}}(\mathrm{t})$ as input to predict the physical KPMs:

$$[T_{CB}, q''_{CB}] = \mathrm{g_{FSM}}\big(Q(t)_{update}, X_{FSM}\big), \quad \text{(Eq. 18)}$$

This whole coupled environment can be functionally represented as a function that consists of nested functional models of the three modules. Let, $\mathrm{g}_{coupled}$ be the functional representation of the coupled environment. Then,

$$[T_{CB}, q''_{CB}] = \mathrm{g}_{coupled}(\mathrm{g_{FSM}}, \mathrm{g_{HPM}}, \mathrm{g_{AM}}, X_{FSM}, X_{HPM}, X_{AM}\ ), \quad \text{(Eq. 19)}$$

where $\mathrm{T_{CB}}$ and $\mathrm{q''_{CB}}$ are the physical KPMs of the target cable (i.e., maximum temperature and heat flux of the cable jacket).

2.2. Advancing the Probabilistic Validation (PV) Methodology

The PV methodology lies at the core of the approach used to develop a generic methodology, evaluating automation trustworthiness. This section starts with a brief introduction to the PV methodology, previously developed by some of the authors of this work [8, 9]. This is followed by the explanation on how PV methodology could be leveraged for automation trustworthiness evaluation. Finally, the methodological advancements needed to operationalize PV for automation trustworthiness evaluation are going to be highlighted. These advancements are made keeping in mind the generic bidirectional interactions within a coupled human-physics-automation model and then assessed over a case study, involving AI-based automated firewatch, in this work.

PV methodology has been originally developed to facilitate the validity evaluation of advanced simulation models that are used for Probabilistic Risk Assessment (PRA) in support of risk-informed decision-making and regulation. This PV methodology has been found particularly useful especially when validation data are not sufficiently available. It advances the scientific usage of uncertainty and acceptability criteria to facilitate the validity evaluation for simulation predictions. PV methodology runs a comprehensive uncertainty analysis over a simulation model, under which all the uncertainty sources associated to the model inputs, the model itself and the numerical approximations introduced during the development process of that simulation model are identified and characterized. The characterized uncertainties are then propagated under bottom-up approach [8, 9]. That is, all the uncertainty sources at the level of simulation model's input are propagated up to their desired level. The desired level could be either at the level of simulation model's output where simulation model would be making predictions or at plant’s risk metric level, depending upon where the predefined acceptability criteria is available. Then total uncertainty would be quantified at the desired level. Uncertainty sources, associated to a simulation model, which are being referred to as uncertainty parameters later in this section, can be of two types: (i) epistemic uncertainty that arises from a lack of knowledge regarding the true values of the predicted quantities; and (ii) aleatory uncertainty rooted from inherent stochasticity of the physical phenomena underlying the quantities of interest. In the PV methodology, the validity of a simulation prediction used for PRA is determined by firstly estimating the magnitude of epistemic uncertainty in the simulation prediction and then finding the result of an acceptability evaluation that determines whether the total uncertainty (including both aleatory and epistemic uncertainties) associated with the simulation prediction is acceptable for the specific application of interest (e.g., PRA) [8, 9].

The generic methodology to estimate automation trustworthiness using the mentality of PV borrows the similar two-step process as described above: (1) At first, the degree of trustworthiness of an automation technology is determined by the magnitude of the epistemic uncertainty (i.e., representing the degree of confidence of that automation technology) quantified at the desired level of interest (i.e., at automation output level or plant's risk level). The magnitude of epistemic uncertainty is calculated by a comprehensive uncertainty analysis mentioned above; and (2) the total uncertainty (i.e., the combination of epistemic and aleatory uncertainties) quantified at the desired level is then used for evaluating the acceptability of the corresponding degree of automation trustworthiness by comparing it with a predefined acceptability criterion for a specific application of interest, available at that desired level. This study aims at developing the methodology to evaluate automation trustworthiness as generic as possible addressing a range of autonomous as well as semi-autonomous systems. Within such complex systems, spatiotemporal bidirectional interactions exist between automation, human actions and physical phenomenon. These interactions have been explained inside a coupled human-physics-automation model earlier in this paper along with some examples in subsection 2.1.1.4.

In order to operationalize the proposed methodology to evaluate automation trustworthiness for a coupled human-automation-physics model, the underlying interactions within that coupled model should be analyzed as realistically as possible. For example, if the signal provided by an automation system to a MCR operator gets delayed due to some reasons, its influence should be reflected on requisite human actions that need to be taken by the operator following that signal. Hence, dependencies between different interactions within a coupled human-automation-physics model need to be assessed correctly and realistically. These interactions within a coupled human-automation-physics model are represented by a list of uncertain parameters (Table 3 in subsection 2.2.1). The current scope of PV methodology [8, 9] does not account for such dependencies between uncertain parameters and hence needs to be updated for the methodology to evaluate automation trustworthiness. The two key advancements made to the existing PV methodology before leveraging it to estimate automation trustworthiness are: (a) Development of a Phenomenon Identification and Ranking Table (PIRT) for automation technology (Step 1 of the methodology), identifying list of uncertain parameters associated to that technology; and (b) Evaluating correlation coefficients between the dependent uncertain parameters associated to a coupled human-automation-physics model to correctly sample (Step 4 of the methodology) and propagate (Step 6 of the methodology) the uncertain parameters to their desired level of interest. The methodology to evaluate automation trustworthiness has been explained by a list of steps in subsection 2.2.1. Each of the steps has been explained for a generic human-automation-physics coupled model followed by a brief explanation tailored for the case study on AI-based automated firewatch. While the methodology to evaluate automation trustworthiness has been developed, the case study is still under process and the detailed results are going to be added in the future publications of this paper.

2.2.1. The methodology to evaluate automation trustworthiness

Methodology to evaluate automation trustworthiness using the mentality of PV follows the steps explained below. For the steps (1,4 and 6) that have been advanced to estimate automation trustworthiness are carrying more details, focusing on problem definitions and how those gaps could be addressed for the present case study. Details for the rest of the steps could be found in (Bui et al., 2023) [8, 9]. Each step has been explained for a generic human-automation-physics coupled model followed by its application over the case study focused on AI-based automated fire classifier. This case study involves a multi-level and a single-model system (figure 7 in step 2). The different hierarchical levels are denoted by i, models are denoted by j and different model-forms are denoted by k. The case study is regarded to have a single-model system because the three different modules of AM, FSM and HPM are coupled together into a single model as shown in figure 7. So, if $M_{i,j,k}$ represents a module in figure 7 over which the first 10 steps of the methodology need to be applied, then starting from the bottom (input level) to the top (system level), the modules of AM, FSM and HPM could be represented as $AM_{1,1,2}$, $FSM_{1,1,1}$ and $HPM_{1,1,1}$ respectively (figure 7). More details about these subscripts are going to be explained in step 6 and step 10.

Step#1: Developing theoretical causal frameworks to identify uncertain parameters influencing the coupled human-automation-physics model's output: This step guides the process of identifying important causal factors and their paths of influence. Such causal factors are in fact the sources of uncertainties that could influence the outputs of a coupled human-automation-physics model. The theoretical causal model developed in this step helps ensure that the uncertainty quantification for the coupled model's outputs (Step 10 in this methodology) will not be missing any important sources of uncertainties. In this work, the existence of the three modules of human, automation and physics within the coupled model, has motivated the development of separate causal models for AM, FSM and HPM. This is because these causal models would be carrying the footprint of all the important activities performed during the development processes of AM, FSM and HPM. For instance, the development process of AM should include the hardware and software development life cycle processes. In this methodology, these causal influencing factors or sources of uncertainties would be referred to as uncertain parameters associated to a coupled human-automation-physics model.

Several frameworks such as Bayesian Belief Network (BBN) or Phenomenon Identification and Ranking Table (PIRT) could be used to develop such theoretical causal models. Previously PV methodology has been operationalized for FSM (that used CFAST) only and hence the theoretical causal model has been developed only for the FSM, without consideration of any dependencies between uncertain parameters [8, 9]. In this work, FSM inside the coupled model also utilized CFAST, so the existing causal model developed by (Bui et al., 2023) has been borrowed [8, 9]. For HPM, a list of important sources of uncertainties relevant to our case study has been adopted from a previous work of some of the authors [18]. Finally for AM, a PIRT has been established as one of the key advancements made to the existing PV methodology. It should be noted that amongst different theoretical causal frameworks as described above, PIRT offers one of the most systematic ways to identify important sources of uncertainties. Hence, PIRT can be established for FSM and HPM as well, in the future publications of this work.

An example of a PIRT table developed for AM could be seen in Table 2. This PIRT table has tailored for AI-based automated fire classifier, keeping it consistent with the case study. To develop PIRT, a questionnaire has been generated enlisting important phenomenon that could influence the AI-based fire classifier's performance (First column). The questionnaire has been distributed amongst experts from the fields of machine learning, fire detection and analysis and NPP operations. In this questionnaire, experts were asked to rank these phenomena based on their importance in influencing the AI-based fire classifier's performance (Second column) and based on the level of knowledge experts have regarding those phenomena (Third column). This is done in accordance with Nuclear Regulatory Commission (NRC) guidelines on developing PIRT table [19]. The expert panel is comprised of six experts and so far, only two experts (E-A and E-B) have reported their feedback that resulted in table 2. Complete results from all the experts will be shared in the future publication of this work.

Table 2: PIRT Table for AI-based Fire Classifier on the Scale of 1 to 5

Phenomenon	Importance		Current Understanding		Degree of Expert's Confidence	
	E-A	E-B	E-A	E-B	E-A	E-B
Training Data Quality	3	5	2	3	5	1
Model Architecture	4	2	3	4	5	5
Model Training	4	2	3	2	3	3
Model Hyperparameters	3	Not sure	2	3	Not sure	Not sure
Feature Extraction & Selection	3	4	2	2	3	3
Noise and Anomalies	2	Not sure	3	4	Not sure	Not sure
Regularization Techniques	4	4	4	4	5	3
Model Interpretability	3	1	2	3	5	1
Operational Environment	2	5	3	3	5	3

As a result of step 1, all the important uncertain parameters associated to a coupled human-automation-physics model were identified. In this methodology, for the sake of understanding, the identified uncertain parameters are categorized into input uncertain parameters and intermediate uncertain parameters. Input uncertain parameters are those parameters that serve as an input to FSM, AM or HPM. On the other hand, intermediate uncertain parameters are the one that come into picture while running FSM, AM or HPM with their associated input parameters. The list of these uncertain input and intermediate parameters for the present case study could be found in Table 3 and Table 4 under step 2 and step 4 of this proposed methodology, respectively.

Step#2: Approximate characterization and qualitative screening of input uncertain parameters influencing the outputs of human-automation-physics coupled model: Step 2 focuses on two key tasks: (I) Characterization of input uncertain parameters identified in step 1; and (II) Identifying correlated uncertain parameters from uncorrelated uncertain parameters followed by estimating correlation coefficients between correlated parameters. The characterization of input uncertain parameters involves two steps: 1) Selection of the choice of distribution functions, modeling input uncertainties, using tests such as probability plotting or goodness of fit tests (like K-S tests); and 2) Parameter estimation associated to each of these distribution functions using Maximum Likelihood Estimate (MLE) or Bayesian techniques. Following approximate characterization, qualitative screening of all the input uncertainties needs to be done to distinguish important parameters from unimportant ones, qualitatively. Expert judgement, literature review findings and insights coming from the coupled automation-physics-human model developed in subsection 2.1.1 could be leveraged to conduct this qualitative screening process. For the present case study, focused on AI-based automated firewatch, the list of input uncertain parameters after qualitative screening is presented in Table 3. Feeding these input parameters to the coupled model would result in intermediate parameters enlisted in Table 4. Each of the intermediate parameters is mentioned with a module (AM, FSM or HPM) responsible for its generation referred to as origin in Table 4.

Table 3: List of Input Uncertain Parameters with their Approximate Characterization for the Case Study

I	Input Uncertain Parameter (Units)	Distribution	Source
1	Maximum Heat Release Rate (KW)	Gamma ($0.271 \le \alpha \le 1.8$, $57.4 \le \beta \le 141$)	NUREG 6850 and NUREG 2233
2	Time to peak HRR (s)	Uniform (38, 1961)	NUREG 2233
3	Steady burning at peak HRR (s)	Uniform (2, 2268)	NUREG 2233
4	Time to decay (s)	Uniform (68, 2581)	NUREG 2233
5	Cable jacket thickness (mm)	Gamma (α = 17.24, β = 0.078)	NUREG/CR-6931
6	Fire location (m)	X = 33, Z= 0.8, Y = Uniform (2, 16.3)	NUREG 1934
7	Visibility constant [10]	Uniform (5, 10)	Bui et al., 2020
8	Time to detect and diagnose the fire signal (s) by the Operator on the MCR computer	Truncated Normal (μ=15, σ=11.7)	(Sakurahara et al., 2020) [18]
9	Time taken by operator to communicate with First Responder (FR) (s)	Truncated Normal (μ=75, σ=35.1)	(Sakurahara et al., 2020) [18]
10	Time taken by FR to arrive and enter fire location, and communicate with operator (s) [20]	Uniform (120, 480)	NUREG 2180
11	Time taken by FR to search for and locate fire source (s)	Truncated Normal (μ=35, σ=19.5)	(Sakurahara et al., 2020) [18]
12	Time to start manual fire suppression by FR using portable fire extinguisher (s)	Truncated Normal (μ=20, σ=7.8)	(Sakurahara et al., 2020) [18]
13	Time taken by operator to receive fire signal from FR to call Fire Brigade (FB) (s)	Truncated Normal (μ=15, σ=11.7)	(Sakurahara et al., 2020) [18]
14	Time taken by the operator to communicate with FB (s)	Truncated Normal (μ=75, σ=35.1)	(Sakurahara et al., 2020) [18]

15	Time taken by FB to gather at the assembly area (s)	Truncated Normal (μ=540, σ=255)	(Sakurahara et al., 2020) [18]
16	Time taken by FB to get pre-plan from MCR Operator (s)	Truncated Normal (μ=45, σ=11.7)	(Sakurahara et al., 2020) [18]
17	Time taken by FB to peer-check equipment (s)	Truncated Normal (μ=45, σ=11.7)	(Sakurahara et al., 2020) [18]
18	Time taken by FB to move to staging area (s)	Truncated Normal (μ=35, σ=19.5)	(Sakurahara et al., 2020) [18]
19	Time taken by FB to arrive at staging area (s)	Truncated Normal (μ=150, σ=23.4)	(Sakurahara et al., 2020) [18]
20	Time taken by FB to lay hose (s)	Truncated Normal (μ=630, σ=164)	(Sakurahara et al., 2020) [18]
21	Time taken by FB to put on breathing apparatus (s)	Truncated Normal (μ=180, σ=46.8)	(Sakurahara et al., 2020) [18]
22	Time taken by FB to search and locate Fire (s)	Truncated Normal (μ=35, σ=19.5)	(Sakurahara et al., 2020) [18]
23	Time taken by FB to prepare the hose and start suppressing the fire (s)	Truncated Normal (μ=180, σ=46.8)	(Sakurahara et al., 2020) [18]

Table 4: List of Intermediate Uncertain Parameters for the Case Study

0	Intermediate Uncertain Parameters	Origin
1	Time evolution of optical density (m^{-1})	FSM with non-suppressed parameters
2	Maximum compartment temperature (°C)	FSM with non-suppressed parameters
3	Time to prompt detection by AI-based AFW (s)	AM with non-suppressed parameters
4	Probability of prompt detection by AI-based AFW	AM with non-suppressed parameters
5	Suppressed maximum Heat Release Rate (KW)	HPM with non-suppressed parameters
6	Suppressed time to peak HRR (s)	HPM with non-suppressed parameters
7	Suppressed steady burning time at peak HRR (s)	HPM with non-suppressed parameters
8	Suppressed time to decay (s)	HPM with non-suppressed parameters
9	Time to suppression (s)	HPM with non-suppressed parameters
10	Suppressed target temperature (°C) and heat flux (KW/m^2)	FSM with suppressed parameters

While proceeding with the next task of step 2 that is estimating correlation coefficients for correlated uncertain parameters, then as explained at the beginning of section 2.2, existence of spatiotemporal and bidirectional interactions between physics, human and automation modules within the coupled model, give rise to dependencies between different input and intermediate uncertain parameters. This comes in contrast to the existing PV methodology in which all the sources of uncertainties were considered to be independent of each other [8, 9]. Before proceeding towards the propagation of these input and intermediate uncertainties in step 6, the correlations between different uncertain parameters need to be determined to account for any unrealistic sampling of these parameters that ultimately may lead to incorrect propagation and quantification of the uncertainties. To distinguish uncorrelated parameters from correlated parameters, Figure 7 shows interactions between HPM, AM and FSM as a coupled model for the case study. This Figure 7 has been generated from the insights coming from Figure 2 in subsection 2.1.1. The solid lines indicate interactions within two modules utilizing intermediate uncertain parameters ranging from O_1 to O_9. The dotted lines represent input uncertain parameters ranging from I_1 to I_{21}. The subscripts of I and O in Figure 7 have been kept in consistent with their numbers in Table 3 and Table 4. This coupled model would operationalize the event tree-based Scenario Based Failure Model (SBFM) as explained in 2.1.1.5, providing system level predictions such as component failure probability, which could serve as an input to PRA.

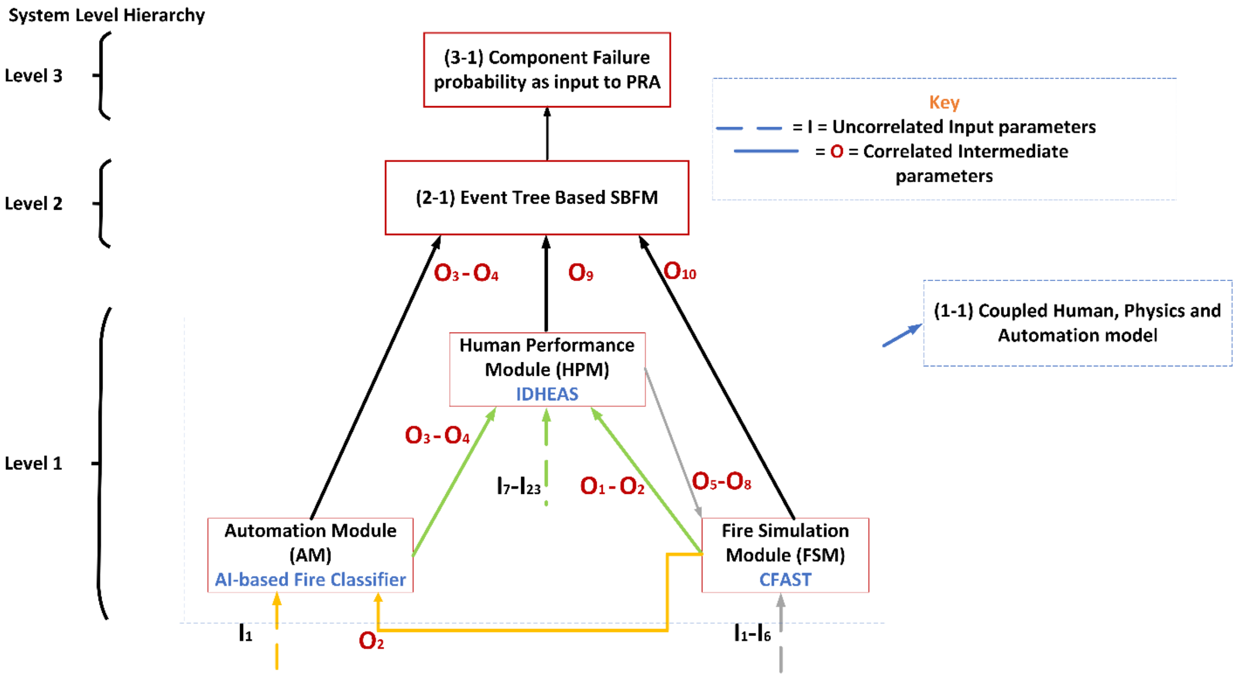
System Level Hierarchy
Level 3
Level 2
Level 1
(3-1) Component Failure probability as input to PRA
(2-1) Event Tree Based SBFM
Key
= I = Uncorrelated Input parameters
= O = Correlated Intermediate parameters
O_3-O_4
O_9
O_{10}
(1-1) Coupled Human, Physics and Automation model
Human Performance Module (HPM)
IDHEAS
O_3-O_4
I_7-I_{23}
O_1-O_2
O_5-O_8
Automation Module (AM)
AI-based Fire Classifier
Fire Simulation Module (FSM)
CFAST
I_1
O_2
I_1-I_6

Figure 7: Interactions within human-automation-physics coupled model leading to correlations and system level hierarchy for the case study.

Due to the nature of interactions within the coupled model, dependencies can exist within input parameters (Table 3) as well as within intermediate parameters (Table 4). For instance, and as shown in Figure 7, there should be a correlation between I_1 "time to receive and detect fire signal" and I_{13}"time to manual fire suppression". In other words, change in the detection time should have an influence on the suppression time that should be reflected while sampling these uncertain parameters. Similarly, talking about intermediate parameters (Figure 7), O_1 to AM represents the CFAST informed optical density of the compartment. The optical density that changes with time triggers the visibility conditions of the compartment and hence should have an influence over the performance of AI-based fire classifier. Similarly, if the parameters associated to HRR curve (I_1- I_4) are greater in size leading to a bigger and more visible flame for longer period of time, it would be easier for the AI-based fire classifier to detect it in less amount of time (O_3) but would be requiring more time for the fire responders to suppress it (O_9). That is, O_3 gets smaller but O_9 would get larger, depending upon the given conditions and the underlying correlation that exist between them. For the present case study, even though some correlations between input uncertain parameters were acknowledged, they have been treated as independent in this work for the sake of simplicity and due to lack of simulation or experimental data at the input level to support their correlation coefficients estimation. However, treatment of correlated uncertain parameters has been executed at the level of intermediate uncertain parameters, mainly because of the availability of simulation data at that level achieved by running the coupled model with input parameters and also because these intermediate parameters (Figure 7) would then be utilized to run the event tree based-SBFM to estimate component failure probability as an input to PRA. The component failure probability gets influenced by the dependency between intermediate parameters making it more critical to be analyzed properly.

Once, all the correlated uncertain parameters are identified a procedure needs to be devised to characterize the dependency among them, providing a pathway to sample these correlated inputs with a consideration of their dependencies to correctly propagate them in step 6. Existing approaches to characterize the dependency among the correlated inputs include: (a) causal model approach in which cause and effect relationships between different parameters need to be developed and illustrated; and (b) developing the correlation coefficients based on available simulation and experimental data. It has been identified that developing causal models require comprehensive knowledge of the problem along with their scientific and theoretical justifications to characterize those causal relationships. For the present case study, availability of a coupled human-automation-physics model can help in creating substantial amount of simulation data and hence could be leveraged to proceed with the approach of generating correlation coefficients for the correlated inputs.

The two most common approaches for developing correlation coefficients include the Pearson correlation and the Spearman Rank Correlation Coefficient (RCC) methods [21]. The Pearson correlation matrix, which is one of the parametric methods, requires the assumption that all the input variables should be represented by a normal distribution, while the Spearman RCC matrix is one of the non-parametric methods that gives relaxation over the choice of distribution functions for the correlated inputs [22]. In this work, Spearman's RCC method is adopted due to the fact that the parameters characterized in this case study are modeled by different kinds of distribution functions as indicated by Table 3. Equation 20 is used to calculate the Spearman RCC for each combination of the two correlated inputs x_i and x_j [23].

$$R_{x_i x_j} = \frac{\sum_{t=1}^{n}(R_i^t - \frac{(1+n)}{2})(R_j^t - \frac{(1+n)}{2})}{\sqrt{\sum_{t=1}^{n}(R_i^t - \frac{(1+n)}{2})^2 (R_j^t - \frac{(1+n)}{2})^2}} \quad \text{(Eq. 20)}$$

Where R_i^t and R_j^t are the ranks among n data points from low to high ($t = 1, \ldots, n$) associated with x_i and x_j, respectively. Unlike the Pearson correlation matrix method, the Spearman RCC method utilizes the ranking of the data points instead of relying on the raw data. If the same rank is identified, the averaged value of the rank will be assigned for each data point. The values of $R_{x_i x_j}$ range from -1 to 1, with -1 showing a perfect negative (increase in x_i would result in x_j to decrease) monotonic relationship and 1 showing a perfect positive (if x_i increases; x_j increases as well) monotonic relationship. A Spearman Rank Correlation matrix (C_R) associated with d correlated inputs is shown in equation 21, where each element of C_R is calculated using equation 5. It should be noted that whenever $i = j$ inside C_R, that entry inside the RCC matrix would be equal to 1, indicating that no dependency exists between a similar correlated input uncertainty source. The developed RCC matrix would help in correctly sampling these corelated uncertain parameters. The RCC matrix for the present case study is still under process and results would be shared in the future publication of this work.

$$C_R = \begin{bmatrix} 1 & R_{x_1 x_2} & \cdots & R_{x_1 x_d} \\ R_{x_2 x_1} & 1 & \cdots & R_{x_1 x_2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{x_d x_1} & R_{x_d x_2} & \cdots & 1 \end{bmatrix} \quad \text{(Eq. 21)}$$

Finally, as explained above, equations 20 and 21 help in establishing correlations between different intermediate uncertain parameters. However, these correlations are only going to signify statistical relationships between different intermediate parameters by explaining how certain parameters change together. For instance, a coefficient of $R_{O_3 O_9} = 0.9$ shows that if O_3 increases, O_9 should increase as well.

This however does not show any causation between the O_3 and O_9 and it cannot be deduced from $R_{O_3O_9}$ that a change in O_9 is because of O_3 and it could be either coincidental or due to some other intermediate parameters. This motivates in establishing more meaningful relationships between intermediate parameters by looking into their underlying causation and not just correlation. In this case study, after quantifying the RCC matrix for intermediate parameters, the coupled human-automation-physics model would be run under controlled settings (designed to study the influence of intermediate parameters with each other) to evaluate causal relationships between different parameters. Detailed approach and results are going to be added in the future publication.

Step#3: Quantitative screening of input uncertainties associated to a coupled model: In this step, all the uncertain parameters retained in step 2 would be ranked based on their influence on coupled model's output. This step would be done quantitatively using Extended Morris Elementary Effect analysis [24]. Only those uncertain parameters identified to be important under this quantitative screening method, would be propagated through the human-physics-automation coupled model to estimate total uncertainty during next steps of the methodology. The details on how quantitative screening can be done using Morris analysis could be found in (Bui et al., 2023) [8, 9].

Step#4: Detailed characterization of input uncertainties and estimation of correlation coefficients for intermediate uncertain parameters associated to the coupled model: In this step at first, the retained uncertain parameters following step 3 would undergo with their detailed characterization using more available data [8, 9]. This is followed by the classification of these uncertain parameters into epistemic or aleatory uncertainties to separately treat the two kinds of uncertainties during step 6. Separate treatment of the aleatory and epistemic would help in analyzing the influence of each of these uncertainty sources over degree of automation trustworthiness to get a better insight on where to focus in order to improve the trustworthiness of that system, during step 12 of this methodology. For the case study, the uncertain parameters identified to be epistemic and aleatory are enlisted in Table 5. Parameters are regarded as pure aleatory if their uncertainty is coming from inherent variability such as timings associated to human actions. On the other hand, parameters showing uncertainty due to lack of information are referred to as pure epistemic. For instance, due to lack of information regarding where fire is located inside CSR, the parameter of fire location is regarded as pure epistemic. Finally, the parameters exhibiting uncertainty due to the said reasons above at the same time, are classified as mixed aleatory and epistemic such as maximum HRR, in this case study.

Table 5: Detailed Characterization of Uncertain Parameters

Uncertain Parameters	Type	Uncertain Parameters	Type
Maximum Heat Release Rate (KW)	Mixed aleatory and epistemic	Time taken by operator to receive fire signal from FR to call Fire Brigade (FB) (s)	Pure aleatory
Time to peak HRR (s)	Pure aleatory	Time taken by the operator to communicate with FB (s)	Pure aleatory
Steady burning at peak HRR (s)	Pure aleatory	Time taken by FB to gather at the assembly area	Pure aleatory
Time to decay (s)	Pure aleatory	Time taken by FB to get pre-plan from MCR Operator (s)	Pure aleatory
Cable jacket thickness (mm)	Pure epistemic	Time taken by FB to peer-check equipment (s)	Pure aleatory
Fire location (m)	Pure epistemic	Time taken by FB to move to staging area (s)	Pure aleatory

Time to detect and diagnose the fire signal (s) by the Operator on the MCR computer	Pure aleatory	Time taken by FB to arrive at staging area (s)	Pure aleatory
Time taken by operator to communicate with First Responder (FR) (s)	Pure aleatory	Time taken by FB to lay hose (s)	Pure aleatory
Time taken by FR to arrive and enter fire location, and communicate with operator (s)	Pure aleatory	Time taken by FB to put on breathing apparatus (s)	Pure aleatory
Time taken by FR to search for and locate fire source (s)	Pure aleatory	Time taken by FB to search and locate Fire (s)	Pure aleatory
Time to start manual fire suppression by FR using portable fire extinguisher (s)	Pure aleatory	Time taken by FB to prepare the hose and start suppressing the fire (s)	Pure aleatory
Visibility constant	Pure epistemic		

Step#5: Characterization of uncertainties associated to numerical approximation: After the detailed characterization of uncertain parameters and estimating C_R for correlated intermediate parameters, additional uncertainties coming from numerical errors committed during the development life cycle of the coupled model, would be characterized in this step. For example, differential equation-based models like CFAST offer no exact solutions for practical problems and give approximate solutions with some underlying numerical approximations. In general, these numerical errors may include discretization errors, iterative convergence errors, round-off errors, and errors due to computer programing mistakes. However, care should be taken that code verification process needs to be done before step 3 to avoid propagation of coding errors, while doing quantitative screening process. No modifications have been made on this step in the current methodology and more details for this step could be seen in (Bui et al., 2023) [8, 9].

Step#6: Propagation of unscreened uncertain parameters through the coupled model: All the uncertain parameters retained till step 4 would then be propagated from the level of input to the coupled model's output. Uncertainty propagation could be done with various techniques, but the double loop Monte Carlo technique would be used in this work to separately handle the aleatory and epistemic uncertainties identified in step 4 [8, 9]. The propagation is done in a nested iterative fashion. At first, all the uncertain parameters identified to be epistemic in step 4 are sampled on the outer loop and then for each sampled set of epistemic uncertain parameters, aleatory uncertain parameters are sampled on the inner loop. Key steps of the computational procedure for uncertainty propagation using double loop MC simulation include:

1. Repeat the steps a, b and c N_{Ot} times, the number representing the sample size of the outer loop.
    a. Get a sampled set of all epistemic uncertain parameters using Latin Hyper Cube (LHC) sampling. This sampled set would be represented by $e^{(Ot)} = [e^{(1)}, e^{(2)}, \dots, e^{(nX_e)}]$, where Ot represents outer loop and nX_e represents the dimension of the sampled set equal to the total number of epistemic uncertain parameters. For the present case study $nX_e = 3$. For each $e^{(Ot)}$ of the total sets equal to N_{Ot}, conduct sub-steps (i) and (ii), N_{in} times, which is a number representing the sample size of the inner loop.
        i. Get a sampled set of all aleatory uncertain parameters using Latin Hyper Cube (LHC) sampling. This sampled set would be represented by $a^{(in)} = [a^{(1)}, a^{(2)}, \dots, a^{(nX_a)}]$, where in represents the inner loop and nX_e represents the dimension of the sampled set equal to the total number of aleatory uncertain parameters. For the present case study $nX_a = 21$.
        ii. With the help of Raven run the human-physics-automation coupled model using the sampled sets $e^{(Ot)}$ and $a^{(in)}$. This is done as per the procedure explained in

subsection 2.1.1.3. This results into a point estimate for the Key Performance Measure (KPM) of interest, $Y_{i,j,k}(in, Ot) = [AM_{1,1,2}, HPM_{1,1,1}, FSM_{1,1,1}]$. The KPMs of this case study are target maximum temperature and heat flux as shown in figure 2.

b. Getting a point estimate in sub-step (ii) and then running Raven for N_{in} times as mentioned in sub-step (a), a total N_{in} point estimates are generated that are used to generate an empirical CDF for the KPM of interest. The empirical CDF is represented by $F_{Y_{i,j,k}}(N_{in}, Ot)$.

2. Run Raven for N_{Ot} times as mentioned in step (1), a total of N_{Ot} empirical CDFs would be generated. This family of the empirical CDF curves of the KPM of interest would then be used to generate a p-box for that KPM, denoted as $[\bar{F}_{Y_{i,j,k}}, \underline{F}_{Y_{i,j,k}}]$.

It should be noted that for those uncertain parameters that are identified to be correlated in step 2, there is need to convert their uncorrelated samples (generated from normal LHC sampling as described above) to correlated samples. Several approaches such as Iman-Corner procedure could be used to do this conversion [25, 26]. The details and results of this procedure will be added to the future publication of this work.

Step#7: Characterization of model uncertainty associated with the coupled model: In this step, the uncertainties added during the approximations or assumptions made during the model development process of a human-automation-physics coupled model would be estimated. This could be done by using several approaches proposed in the original PV methodology [8, 9], depending upon the availability of validation data. In these approaches, model performance is compared with validation data at the same level to characterize model form uncertainties associated to that model. In the existing PV methodology, if validation data exists and is found to be enclosed within the application domain, then data-driven methodology could be used to characterize the model uncertainty. However, if no validation data exist, then a causal model (similar to the one in step 1 but more specific to factors associated to computational model development rather than input factors to the model) needs to be made, and with a bottom-up approach, model form uncertainty could be characterized.

In the present case study, data driven approach would be utilized to characterize model uncertainty of the coupled human-physics-automation model. To separately characterize the model uncertainties associated to FSM, AM and HPM, existing validation data for CFAST would be leveraged [8, 9]. Similar could be done for HPM by arranging validation data for IDHEAS. In this work, the model uncertainty associated to the AI-based fire classifier needs to be characterized. This is done by comparing the outputs coming from the developed AI-based fire classifier with another AI-based fire and smoke classifier. The AI-based fire and smoke classifier being more realistic in nature would serve as a surrogate reality model as a substitute to actual validation data. This is particularly useful for the cases, where getting actual data coming from the tests might not be feasible.

Step#8: Estimating total uncertainty associated with the coupled model: In this step, the total uncertainty coming from input and intermediate uncertain parameters and coupled model's uncertainty would be quantified by aggregating results from step# 6 and step# 7, just like done in the existing work [8, 9].

Step#9: Bayesian updating the uncertainty associated with the coupled model: If additional empirical data associated with the coupled model is available, this step conducts Bayesian updating to maximize the use of available empirical data and update the total uncertainty associated with the model predictions. This step serves two goals: i) It accounts for the cumulative effect of sources of uncertainty that have not been addressed in the previous steps (e.g., uncertainty due to errors in the screening processes); and ii) When validation domain associated to different models are not available and step 7 could not be feasible [8, 9], step 9 provides an alternative solution to model uncertainty quantification. This is done by Bayesian updating the uncertainty associated with the model response/prediction (obtained in Step 6) with available empirical data (if any) in Step 9. For the present case study, this step would be applied to the coupled human-physics-automation model with the similar way as proposed by (Bui et al., 2023) [8, 9].

Step#10: Aggregating results associated with multiple model forms to estimate uncertainty associated to the coupled model: In this step, the uncertainty associated to the choice of model being used would be considered. For the present case study, no multi model form uncertainty would be considered for FSM and HPM, while the choice of picking up two different models such as CNN or Fully Connected Neural Network (FCNN) to develop an AI-based fire classifier for AM, would lead to its multi model form uncertainty that is going to be aggregated in this step.

Step#11: Evaluating the acceptability of automation system's trustworthiness from some predefined acceptability criteria: This step helps to evaluate the acceptability of the current degree of automation trustworthiness. Evaluating the acceptability of the current degree of automation trustworthiness for a specific application condition is done by comparing the total uncertainty (epistemic and aleatory) associated with the automation system against predefined acceptability criteria, at the desired level (coupled model's output level or plant's risk metric level, depending upon where that acceptability criteria would be available) of system hierarchy. The automation can then be considered "sufficiently trustworthy" for that specific application condition if the total uncertainty satisfies the acceptability criteria.

Step#12: Improving the automation system's trustworthiness with the insights coming from GIM rankings of all the sources of uncertainties associated to the coupled model: The last step is aimed to improve the degree of automation trustworthiness up to a level where the automation would be sufficiently trustworthy and ready for deployment at NPPs. This is done by equipping methodology to evaluate automation trustworthiness with advanced global importance ranking analyses to help identify factors (among those identified in step 4) that contribute the most to the total uncertainty in the coupled model's outputs (or the total uncertainty in the plant risk estimate). The importance ranking results will inform decision-makers regarding how to prioritize their resources most efficiently for improving the degree of automation trustworthiness.

3. Concluding Remarks and Future Work

In this paper, a new methodology estimating trustworthiness using advanced uncertainty analysis approach has been proposed. It develops a systematic and scientifically justifiable link between degree of trustworthiness of automation system and epistemic uncertainty associated to the automation system's output. All the sources of uncertainties associated to the underlying simulation modules of automation, fire simulation and human performance were first identified, characterized and propagated under bottom-up approach. The epistemic uncertainty quantified at the level of human-automation-physics coupled system's output, would be regarded as its degree of trustworthiness. To propagate these uncertainties, a computational platform using I-PRA framework has been utilized. Under this platform, an AM using AI-based fire classifier; a HPM using IDHEAS; and an FSM using CFAST have been developed as I-PRA underlying simulation modules. A methodology has been developed to model spatiotemporal bidirectional interactions between these modules. Under this coupled model, RAVEN tool is used to generate an environment for sampling, with samples coming from the LHC sampling technique. For each sampled set of variables, the coupled model has been run multiple times. Aleatory and epistemic uncertainties were separately treated and propagated using double loop Monte Carlo sampling. The two loops would ensure generating a p-box for the KPM of interest, with the total epistemic uncertainty representing the width of that p-box. The quantified epistemic uncertainty of the desired KPM (e.g., target's maximum temperature) at a specific level of interest (e.g., simulation model's output or plant risk metric) would be regarded as the degree of trustworthiness of the coupled simulation model at that level. The case study used in this paper involves a transient fire scenario in an open trash (Appendix E of NUREG 1934) and focused on AI-based fire classifier as the AM. This paper only explains the methodology used to evaluate automation trustworthiness. Demonstrating the feasibility of the proposed methodology calls for its application over a case study, which is still under process. Complete results of the case study are planned to be added in the future publications of this work.

For future considerations, a correlation matrix needs to be developed explaining the correlations between different uncertain parameters for the coupled human-physics-automation model. Once trustworthiness would be quantified, trustworthiness acceptability evaluation should be added to decide the deployment of automation technologies. Global Importance Measure (GIM) method should also be added to rank significant sources of uncertainty at the level of underlying simulation models, facilitating in improving the degree of trustworthiness of automation systems.

References

1. Muhammad Hammad Khalid, H.B., Pegah Farshadmanesh, Ahmad Al Rashdan and Zahra Mohaghegh *Automation Trustworthiness in Nuclear Power Plants: A Literature Review.* Abstract accepted to the IAPSAM Topical Conference on Artificial Intelligence and Risk Analysis for Probabilistic Safety/Security Assessment & Management, 2023.
2. Wang, L., *Development and Application of Data Coverage Assessment for NAMAC Trustworthiness.* 2022.
3. Kobayashi, K., et al., *Explainable, Interpretable & Trustworthy AI for Intelligent Digital Twin: Case Study on Remaining Useful Life.* arXiv preprint arXiv:2301.06676, 2023.
4. Smidts, C., X. Diao, and P.K. Vaddi, *Next-Generation Architecture and Autonomous Cyber-Defense.* Industrial Control Systems Security and Resiliency: Practice and Theory, 2019: p. 203-234.
5. Sun, B., et al., *A novel concept and assessment method for trustworthiness of prognostics.* Advances in Mechanical Engineering, 2016. **8**(3): p. 1687814016638807.
6. Sakurahara, T., et al., *An integrated methodology for spatio-temporal incorporation of underlying failure mechanisms into fire probabilistic risk assessment of nuclear power plants.* Reliability Engineering & System Safety, 2018. **169**: p. 242-257.
7. Bui, H., et al., *An algorithm for enhancing spatiotemporal resolution of probabilistic risk assessment to address emergent safety concerns in nuclear power plants.* Reliability Engineering & System Safety, 2019. **185**: p. 405-428.
8. Bui, H., et al., *Probabilistic Validation: Theoretical Foundation and Methodological Platform.* ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering, 2023. **9**(2): p. 021204.
9. Bui, H., et al., *Probabilistic Validation: Computational Platform and Application to Fire Pra of Nuclear Power Plants.* ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg, 2023: p. 1-42.
10. Bui, H., et al., *Spatiotemporal integration of an agent-based first responder performance model with a fire hazard propagation model for probabilistic risk assessment of nuclear power plants.* ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering, 2020. **6**(1): p. 011011.
11. Peacock, R.D., P.A. Reneke, and G.P. Forney, *CFAST–consolidated model of fire growth and smoke transport (version 7) volume 2: user's guide.* NIST Technical Note 1889v2, 2017.
12. *EPRI/NRC-RES Fire PRA Methodology for Nuclear Power Facilities Detailed Methodology. Electric Power Research Institute (EPRI), Palo Alto, CA, and U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research (RES), Rockville, MD: 2005, EPRI TR-1011989 and NUREG/CR-6850.* **2**.
13. Mohaghegh, Z., et al. *Risk-informed resolution of generic safety issue 191*. in *International Topical Meeting on Probabilistic Safety Assessment and Analysis 2013, PSA 2013*. 2013.

14. Sakurahara, T., et al., *Methodological and practical comparison of integrated probabilistic risk assessment (I-PRA) with the existing fire PRA of nuclear power plants.* Nuclear technology, 2018. **204**(3): p. 354-377.

15. Juliani, A., et al., *Unity: A general platform for intelligent agents.* arXiv preprint arXiv:1809.02627, 2018.

16. Melly, N., *Methodology for Modeling Transient Fires in Nuclear Power Plant Fire Probabilistic Risk Assessment.* NUREG-2233 and EPRI 3002016054, USNRC, 2020.

17. Salley, M. and R. Wachowiak, *Nuclear power plant fire modeling analysis guidelines (NPP FIRE MAG).* US Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, Washington, DC, and Electric Power Research Institute, Palo Alto, CA, Report No. NUREG-1934 and EPRI, 2012. **1023259**.

18. Sakurahara, T., Z. Mohaghegh, and E. Kee, *Human Reliability Analysis-Based Method for Manual Fire Suppression Analysis in an Integrated Probabilistic Risk Assessment.* ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering, 2020. **6**(1): p. 011010.

19. Goodson, C.E. and K.J. Geelhood, *Degradation and Failure Phenomena of Accident Tolerant Fuel Concepts: FeCrAl Alloy Cladding*. 2020, Pacific Northwest National Laboratory (PNNL), Richland, WA (United States).

20. Taylor, G., *Determining the Effectiveness, Limitations, and Operator Response for Very Early Warning Fire Detection Systems in Nuclear Facilities (DELORES-VEWFIRE)*. 2015: US Nuclear Regulatory Commission, Office of Nuclear Regulatory Research.

21. Kendall, M.G., *Rank correlation methods.* 1948.

22. Hauke, J. and T. Kossowski, *Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data.* Quaestiones geographicae, 2011. **30**(2): p. 87-93.

23. Zwillinger, D. and S. Kokoska, *CRC standard probability and statistics tables and formulae*. 1999: Crc Press.

24. Ge, Q. and M. Menendez, *Extending Morris method for qualitative global sensitivity analysis of models with dependent inputs.* Reliability Engineering & System Safety, 2017. **162**: p. 28-39.

25. Helton, J.C. and F.J. Davis, *Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems.* Reliability Engineering & System Safety, 2003. **81**(1): p. 23-69.

26. Iman, R.L. and W.-J. Conover, *A distribution-free approach to inducing rank correlation among input variables.* Communications in Statistics-Simulation and Computation, 1982. **11**(3): p. 311-334.